# Fusion d'informations pour la compréhension de scènes

Philippe Xu[1,2] Franck Davoine[2] Jean-Baptiste Bordes[1]
Thierry Denœux[1]

[1]CNRS UMR 7253, Heudiasyc
Université de Technologie de Compiègne
BP 20529, 60205 Compiègne Cedex, France

[2]LIAMA, CNRS
Key Lab of Machine Perception (MOE)
Peking University, Pékin, R.P. Chine

**Résumé**

Cet article traite du problème de la compréhension de scènes routières pour des systèmes d'aide à la conduite. Afin de pouvoir reconnaître le grand nombre d'objets pouvant être présents dans la scène, plusieurs capteurs et algorithmes de classification doivent être utilisés. L'approche proposée est fondée sur la représentation de toutes les informations disponibles au niveau d'une image sur-segmentée. La principale nouveauté de la méthode est sa capacité à inclure de nouvelles classes d'objets ainsi que de nouveaux capteurs ou méthodes de détection. Plusieurs classes comme le sol, la végétation et le ciel sont considérées, ainsi que trois capteurs différents. L'approche est validée sur des données réelles de scènes routières en milieu urbain.

# 1 Introduction

La compréhension de scènes routières est une tâche complexe qui amène à considérer de nombreuses sous-tâches, allant de la détection d'objets à la localisation en passant par la reconstruction 3D. Tous ces problèmes ont fait l'objet de nombreux travaux de recherche durant les dernières décennies. Malheureusement, chacune de ces tâches est souvent traitée de manière indépendante et isolée, en n'utilisant qu'un ou plusieurs capteurs spécifiques. Afin de pouvoir profiter au mieux de tous les travaux existants, il devient essentiel de pouvoir fusionner les données issues de toutes les sources d'information à disposition.

Plusieurs questions importantes se posent alors. Comment un détecteur de végétation peut-il, par exemple, aider un détecteur de piéton et vice versa ? Comment les données issues d'un capteur LiDAR (*Light Detection And Ranging*), qui ne perçoit qu'un ensemble discret d'impacts réfléchis par des obstacles, peuvent-elles être fusionnées avec celles d'un module de détection de ciel fondé sur une caméra ? Comment inclure de nouveaux capteurs ou de nouvelles classes d'objets ?

Plus généralement, deux buts critiques doivent être atteints. Le premier est de pouvoir combiner les données provenant de plusieurs modules traitant des classes d'objets différentes et d'être assez flexible pour inclure de nouvelles classes. Le deuxième but est de pouvoir représenter, dans un espace commun, les données issues de capteurs pouvant observer l'environnement de manières très différentes.

## 1.1 Travaux antérieurs

Dans le domaine des véhicules intelligents, les caméras et les LiDAR sont les capteurs les plus souvent utilisés pour la perception. Les capteurs LiDAR ont, par exemple, été utilisés pour détecter les structures statiques et les objets en mouvement, notamment dans des contextes de construction de cartes [25, 26] ou de grilles d'occupation [18]. Les caméras ont, quant à elles, été considérées dans un champ d'applications beaucoup plus large. La détection de piétons est l'un des cas les plus étudiés [5]. Des travaux, plus généraux, de classifications multiclasses pour la compréhension de scènes urbaines ont également été menés [7, 14, 8]. Ces derniers utilisent parfois un système stéréoscopique permettant d'avoir une information de profondeur [14] et peuvent notamment servir à détecter les obstacles mais aussi les zones navigables [2].

En ce qui concerne l'aspect fusion, de nombreuses méthodes fondées sur des systèmes multicapteurs utilisent une approche de type « régions d'intérêt ». Le principe est d'utiliser un premier capteur, par exemple un LiDAR [21], pour sélectionner un ensemble de régions candidates, qui seront ensuite analysées plus finement à l'aide d'autres sources d'information. D'autres approches passent par l'estimation de la configuration géométrique de la scène, en calculant, par exemple, le plan du sol ou le point de fuite [15, 11], afin de contraindre la recherche d'objets. Enfin, un autre type de fusion consiste à combiner différentes caractéristiques visuelles [5] et/ou géométriques [7, 14] afin d'avoir un plus grand pouvoir discriminant. Ces méthodes de fusion sont souvent construites autour d'un problème spécifique prédéfini. Il devient alors relativement difficile d'inclure de nouvelles classes d'objets ou caractéristiques. De même, l'ajout d'un nouveau capteur ou module de traitement est difficilement envisageable. En ef-

fet, il devient nécessaire d'entraîner entièrement le système à chaque fois qu'une nouvelle source d'information doit être considérée.

### 1.2 Contributions

Dans cet article, nous nous attachons à construire un système permettant de combiner les sorties de différents modules de traitement sans contrainte sur leur tâche spécifique ni les capteurs dont ils dépendent. Cette flexibilité permet non seulement de pouvoir envisager de nouvelles classes à venir mais aussi de restreindre les classes à analyser. Ainsi, il ne sera pas nécessaire de définir à l'avance une liste exhaustive des objets pouvant apparaître dans la scène, ce qui impliquerait de devoir construire un détecteur pour chacun d'eux. Pour ce faire, la théorie des fonctions de croyance, aussi connue sous le nom de théorie de Dempster-Shafer [23], est utilisée. En particulier, une forme générale de fonction de masse, dans le cadre d'un problème de classification binaire, est proposée. Elle est construite à partir de la distance d'une observation à un modèle et ses paramètres peuvent être optimisés en minimisant une fonction de perte. La combinaison de plusieurs classifieurs binaires définis sur des classes différentes permet finalement d'avoir une classification multiclasse.

Nous montrons également comment combiner les données issues de sources d'information dont les représentations peuvent être de natures différentes. Dans ce but, nous formulons le problème comme celui de l'annotation d'image en utilisant une image sur-segmentée.

La figure 1 montre une vue d'ensemble du système. Plusieurs capteurs, dont une caméra, observent la scène et leurs données de sortie sont traitées par différents modules indépendants. Ces derniers peuvent indépendamment utiliser les données provenant d'un ou plusieurs capteurs. Les résultats de classification de ces modules, qui concernent a priori des classes différentes d'objets, sont tout d'abord projetés dans un espace de décision commun avant d'être fusionnés au niveau d'une image sur-segmentée.

Nous montrerons comment ce système peut être utilisé dans la pratique, en considérant un système comprenant une caméra stéréo et un capteur LiDAR. Plusieurs modules seront décrits, tout d'abord pour la détection du sol, puis pour un problème plus large incluant la végétation et le ciel. Une validation expérimentale est menée sur des données réelles provenant de la base de données KITTI Vision Benchmark Suite [9].

## 2 Annotation d'images sur-segmentées

Comme expliqué en introduction, non seulement il est nécessaire de raisonner avec des classes différentes mais également avec des représentations différentes des données. Étant dans un contexte d'aide à la conduite, le but est d'avertir le conducteur de dangers potentiels. Ainsi, comme l'image acquise par une caméra est proche de ce que perçoit le conducteur, il semble raisonnable de représenter l'information au niveau d'une image. Plus précisément, la tâche est d'annoter une image, à savoir attribuer une classe à chaque pixel de l'image.

Raisonner au niveau du pixel est souvent trop local et difficile, tandis que raisonner au niveau des objets (*e.g.*, en utilisant des boîtes englobantes) est inadapté pour certaines classes d'objets comme la route. Nous avons choisi un

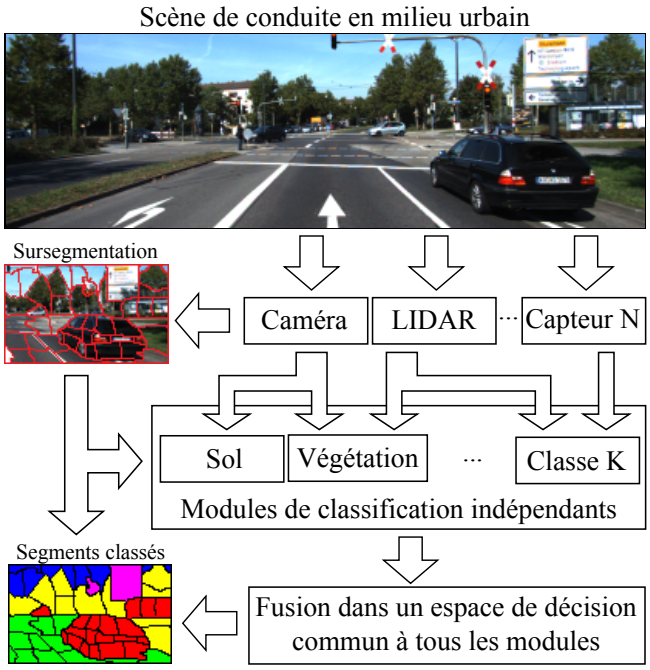



FIGURE 1 – Vue d'ensemble du schéma de fusion. $K$ modules indépendants traitent les données fournies par $N$ capteurs observant la scène, dont une caméra. Les résultats de classification sont ensuite fusionnés dans un espace de décision commun à tous les modules au niveau d'une image sur-segmentée.

(a) (b)

FIGURE 2 – (a) Sur-segmentation obtenue par la méthode de Felzenszwalb et Huttenlocher (2004). (b) Sur-segmentation calculée à partir de l'algorithme SLIC (Achanta *et al.*, 2012).

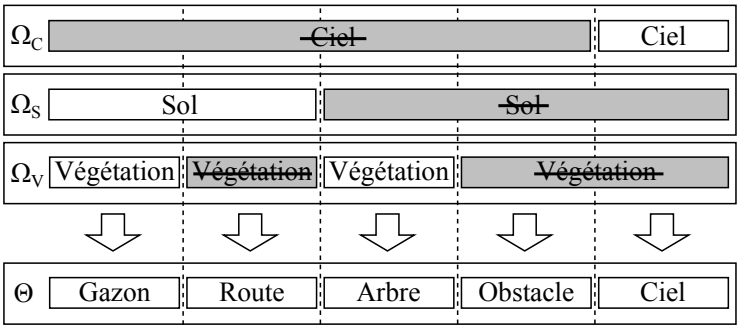



FIGURE 3 – Illustration d'une fusion multiclasse. Les trois premiers blocs représentent trois différentes décompositions du monde. Les blocs gris correspondent aux complémentaires des blocs blancs. En les intersectant, on obtient un raffinement commun. La classe « obstacle » représente, en fait, tout ce qui n'est ni ciel, ni sol, ni végétation. Elle pourrait être, si nécessaire, davantage raffinée pour inclure les piétons ou les voitures.

niveau intermédiaire en sur-segmentant l'image. Cela consiste à regrouper des pixels adjacents pour former un segment, aussi appelé *superpixel*, représentant un objet unique. La méthode de sur-segmentation de Felzenszwalb et Huttenlocher (2004) a été utilisée dans de nombreux travaux portant sur la compréhension de scènes [12, 14, 8]. La figure 2a illustre le résultat obtenu par cette méthode de sur-segmentation. Nous pouvons remarquer que les tailles et les formes des segments sont très hétérogènes. Il est relativement difficile de décrire ce type de sur-segmentation par des notions géométriques comme la hauteur ou la profondeur. D'autres approches [16, 1, 19] permettent d'obtenir une sur-segmentation en forme de grille régulière avec une distribution spatiale et en taille relativement uniforme. L'algorithme SLIC [1] a été choisi en raison de la simplicité de son implémentation. Toutefois, des approches plus complexes utilisant notamment des informations sur la distribution des contours [19] peuvent donner de meilleurs résultats. La figure 2b illustre une sur-segmentation obtenue par la méthode SLIC. Dans notre approche, nous nous limitons à une étude locale des segments. Les relations, notamment d'occultation [13], entre segments adjacents ne sont pas pris en compte.

La tâche commune à tous les modules devient alors de fournir une information sur la classe de chaque segment de l'image. Quelle que soit la représentation en amont, le résultat doit être projeté au niveau de l'image. La théorie des fonctions de croyance permettant de représenter l'ignorance, il n'est pas nécessaire, pour chaque module, de traiter tous les segments de l'image.

# 3 Théorie des fonctions de croyance

Pour montrer l'utilité de la théorie des fonctions de croyance, prenons l'exemple de la fusion des données fournies par un détecteur de sol, un détecteur de ciel et un détecteur de végétation. Comme l'illustre la figure 3, en intersectant ces trois classes initiales, on peut obtenir de nouvelles sous-classes. La classe « sol » peut, par exemple, être divisée en « gazon » et « route » en l'intersectant avec

la classe « végétation ».

Une connaissance spécifique à la classe « sol » ne donne, *a priori*, aucune information sur les classes « gazon » et « route », si ce n'est qu'elles sont toutes les deux aussi plausibles l'une que l'autre. Il est notamment injustifié de distribuer uniformément la connaissance sur la classe « sol » aux classes « gazon » et « route », car une connaissance artificielle quant aux deux nouvelles est créée. Il faut donc pouvoir raisonner sur des ensembles de classes, ce que permet la théorie des fonctions de croyance.

## 3.1 Quelques définitions

Soit $\Omega$ un ensemble de classes mutuellement exclusives, appelé *cadre de discernement*, représentant l'ensemble de toutes les classes d'objets possibles. On appelle *fonction de masse* [23] sur $\Omega$, une fonction $m : 2^\Omega \to [0,1]$ telle que :

$$m(\emptyset) = 0, \qquad \sum_{A \subseteq \Omega} m(A) = 1. \tag{1}$$

Pour une variable $y$ à valeur dans $\Omega$, la croyance relative à son appartenance aux différentes parties de $\Omega$ peut être modélisée par une fonction de masse $m$. Étant donné un sous-ensemble $A$ de $\Omega$, la quantité $m(A)$, qu'on appellera la masse sur $A$, peut être interprétée comme la croyance allouée spécifiquement à l'hypothèse $y \in A$.

Tout sous-ensemble $A \subseteq \Omega$, tel que $m(A) > 0$, est appelé *élément focal* de $m$. On dira que $m$ est une fonction de masse *catégorique* sur $A$, si $A$ est l'unique élément focal de $m$. En particulier, si $A = \Omega$, $m$ représente l'ignorance totale, l'hypothèse $y \in \Omega$ étant supposée toujours vraie. Cette fonction de masse particulière est appelée la fonction de masse vide. On peut remarquer qu'une fonction de masse n'ayant que des singletons comme éléments focaux représente exactement une distribution de probabilité. La théorie des fonctions de croyance est donc une généralisation des probabilités bayésiennes classiques. L'utilisation de la théorie de Dempster-Shafer ne restreint donc pas l'utilisation de méthodes utilisant des représentations probabilistes.

## 3.2 Raffinement d'un cadre de discernement

Partant, d'un cadre de discernement $\Omega$, on peut définir un raffinement $\Theta$ en partitionnant un ou plusieurs éléments de $\Omega$. Sur l'exemple de la figure 3, le cadre de discernement $\Theta$ est un raffinement commun à $\Omega_C$, $\Omega_S$ et $\Omega_V$. Le raffinement de $\Omega$ en $\Theta$ est défini par une application $\rho : 2^\Omega \to 2^\Theta$ telle que :

$$\{\rho(\{\omega\}), \omega \in \Omega\} \subseteq 2^\Theta \text{ est une partition de } \Theta \text{ ;} \tag{2}$$

$$\forall A \in \Omega,\ \rho(A) = \bigcup_{\omega \in A} \rho(\{\omega\}). \tag{3}$$

Par exemple, pour raffiner $\Omega_S = \{Sol, \overline{Sol}\}$ en $\Theta = \{Gazon, Route, Arbre, Obstacle, Ciel\}$, on définit $\rho : 2^\Omega \to 2^\Theta$ par :

$$\rho(\{Sol\}) = \{Gazon, Route\}, \tag{4}$$

$$\rho(\{\overline{Sol}\}) = \{Arbre, Obstacle, Ciel\}. \tag{5}$$

La deuxième propriété (3) sur $\rho$ imposera alors naturellement $\rho(\Omega) = \rho(\Theta)$. On peut alors transformer une fonction de masse $m^\Omega$ définie sur $\Omega$ en une fonction de masse $m^\Theta$ définie sur $\Theta$ en posant pour tout $B \in \Theta$ :

$$m^\Theta(B) \quad = \quad \begin{cases} m^\Omega(A) & \text{si } \exists A \in \Omega,\ B = \rho(A), \\ 0 & \text{sinon.} \end{cases} \tag{6}$$

Une masse initialement assignée à $\{Sol\}$ sera, par exemple, tout simplement transférée sur $\{Gazon, Route\}$ au niveau de $\Theta$.

### 3.3 Affaiblissement

Il est parfois intéressant de pouvoir affaiblir une fonction de masse, notamment lorsqu'on dispose d'une mesure de sa fiabilité. L'affaiblissement d'une fonction de masse $m$ par un facteur $\alpha \in [0,1]$ s'exprime comme :

$$\begin{aligned} {}^\alpha m(A) &= (1-\alpha)m(A), \quad \forall A \subset \Omega, \\ {}^\alpha m(\Omega) &= (1-\alpha)m(\Omega) + \alpha. \end{aligned} \tag{7}$$

Autrement dit, la masse sur tous les éléments focaux est diminuée d'un facteur $1-\alpha$ et le reste est transféré sur l'ignorance.

### 3.4 Combinaison de fonctions de masse

Étant donné un cadre de discernement $\Omega$ et deux fonctions de masse $m_1$, $m_2$, construites à partir de sources d'information indépendantes, elles peuvent être combinées pour former une nouvelle masse $m_{1,2} = m_1 \oplus m_2$, en utilisant la règle de combinaison de Dempster :

$$\begin{aligned} m_{1,2}(\emptyset) &= 0, \\ m_{1,2}(A) &= \frac{1}{1-\kappa} \sum_{B \cap C = A} m_1(B) m_2(C), \end{aligned} \tag{8}$$

avec $\kappa = \sum_{B \cap C = \emptyset} m_1(B) m_2(C)$, qui est une mesure du conflit entre les deux sources d'information. Cette règle de combinaison est associative et commutative, l'ordre dans lequel les sources sont combinées n'a donc pas d'influence sur le résultat final. Pour combiner deux fonctions de masse définies sur des cadres de discernement différents, il suffit de trouver un raffinement commun puis d'utiliser la règle de Dempster.

### 3.5 Prise de décision

Enfin, pour passer d'une fonction de masse à une prise de décision, plusieurs approches sont possibles. La plus répandue consiste à calculer la probabilité pignistique [24] $BetP$ pour tous les éléments $\omega \in \Omega$ :

$$BetP(\omega) \quad = \quad \sum_{A \subseteq \Omega,\ \omega \in A} \frac{m(A)}{|A|}. \tag{9}$$

La masse allouée à un ensemble est uniformément distribuée à tous ses éléments. Puis, le singleton avec la plus grande probabilité pignistique est sélectionné.

Cependant, cette approche n'est pas du tout adaptée dans un contexte où le cadre de discernement peut être changé, en particulier raffiné. Le calcul de la probabilité pignistique utilisant les cardinaux des parties de $\Omega$, il est directement influencé par le découpage du cadre de discernement.

Une autre approche consiste à choisir le singleton avec la plus grande mesure de plausibilité $pl$, qui est définie pour tout $A \subseteq \Omega$ par :

$$pl(A) \quad = \quad \sum_{B \subseteq \Omega,\ B \cap A \neq \emptyset} m(B). \tag{10}$$

Cette approche est beaucoup plus adaptée car la plausibilité accordée à un sous-ensemble donné reste inchangée même si le cadre de discernement est modifié. L'information représentée par la plausibilité est équivalente à celle de la fonction de masse, qui peut être recalculée par :

$$m(A) \quad = \quad \sum_{B \subseteq A} (-1)^{|A|-|B|} \left(1 - pl\left(\overline{B}\right)\right). \tag{11}$$

De plus, comme montré dans [3], étant données $k$ fonctions de plausibilité, il suffit d'un nombre d'opérations en $O(k|\Omega|)$ pour trouver le singleton de plausibilité maximale associé à la combinaison des plausibilités. En effet, la combinaison de Dempster donne pour la plausibilité :

$$\begin{aligned} \forall \omega \in \Omega, \quad pl_{1,2}(\{\omega\}) \quad &= \quad \frac{1}{1-\kappa} pl_1(\{\omega\}) pl_2(\{\omega\}) \\ &\propto \quad pl_1(\{\omega\}) pl_2(\{\omega\}). \end{aligned} \tag{12}$$

La combinaison de Dempster, quant à elle, nécessite, dans le pire des cas, un nombre d'opérations exponentiel en $|\Omega|$. Dans nos travaux, nous utiliserons donc le singleton de plausibilité maximale.

## 3.6 Fonction de masse à partir d'un modèle

Un problème de classification peut souvent être vu comme la recherche d'une correspondance entre un modèle $M$ d'une classe $C$ et une observation $X$ d'un objet $S$ (qui dans notre cas est un segment d'image). À partir d'une mesure $d(X,M)$ entre une observation et un modèle, la classe de l'objet $S$ peut être inférée.

Pour construire une fonction de masse $m$ sur un cadre de discernement $\Omega = \{C,\overline{C}\}$, où $\overline{C}$ comprend tout ce qui n'appartient pas à $C$, nous proposons de définir deux seuils $d^-$ et $d^+$ vérifiant les propriétés suivantes :

$$\begin{array}{lll} a) & \text{Si } d^- > d(X,M) \to 0, & \text{alors } m(\{C\}) \to 1. \\ b) & \text{Si } d^- \leq d(X,M) \leq d^+, & \text{alors } m(\{\Omega\}) \ = 1. \\ c) & \text{Si } d^+ < d(X,M) \to +\infty, & \text{alors } m(\{\overline{C}\}) \to 1. \end{array} \tag{13}$$

Ainsi, quand l'observation $X$ est proche du modèle $M$, la masse mise sur la classe $C$ est proche de 1. Au contraire, si $X$ est très éloignée de $M$, c'est sur $\overline{C}$ que la masse est mise. Enfin, il est parfois intéressant de fixer une zone d'ignorance, qui est vide si $d^- = d^+$, dans laquelle aucune décision ne peut être prise. Si la mesure $d(X,M)$ est bornée, il faut évidemment remplacer, dans les propriétés $a$) et $c$), le 0 et le $+\infty$ par la borne inférieure et supérieure respectivement.

Il est important de noter que, dans certains cas, on ne peut inférer la classe de $S$ que lorsque la valeur de $d(X,M)$ est grande, tandis que rien ne peut être dit dans le cas contraire. Pour gérer cette configuration, il suffit de choisir $d^- = 0$. De même, le cas opposé est traité avec $d^+ = +\infty$.

Une forme générale de fonction de masse satisfaisant les trois propriétés $a$), $b$) et $c$), inspirée de celle proposée par Denœux (1995), est :

$$\begin{aligned} m(\{C\}) &= \begin{cases} e^{-\gamma\left(\frac{d}{d^- - d}\right)^\beta} & \text{si } d < d^- \\ 0 & \text{sinon,} \end{cases} \\ m(\{\overline{C}\}) &= \begin{cases} e^{-\gamma\left(\frac{d^+}{d - d^+}\right)^\beta} & \text{si } d > d^+ \\ 0 & \text{sinon,} \end{cases} \\ m(\Omega) &= 1 - m(\{C\}) - m(\{\overline{C}\}). \end{aligned} \tag{14}$$

Les seuils $d^-$ et $d^+$ définissent les valeurs au-dessous et au-dessus desquelles une décision peut être prise. Le paramètre $\beta \in \{1,2,\ldots\}$, qui peut être arbitrairement fixé à 1 ou 2, comme suggéré dans [4], et $\gamma > 0$ reflètent l'influence de la distance quant à la masse allouée. La fonction de masse obtenue peut, finalement, être affaiblie par un facteur $\alpha$ si nécessaire.

À partir d'une base d'entraînement $\{(X_i,c_i)\}_{1\leq i\leq n}$, où $c_i \in \{C,\overline{C}\}$ est la classe de l'observation $X_i$, les paramètres $d^-$ et $d^+$ peuvent être déterminés par validation croisée en faisant varier leurs valeurs sur une grille. Une fois ces deux paramètres fixés, le paramètre $\gamma$ peut être choisi pour minimiser la fonction de perte suivante :

$$L = \sum_{i=1}^{n} (1 - pl_i(\{c_i\}))^2 + pl_i(\{\overline{c_i}\})^2, \tag{15}$$

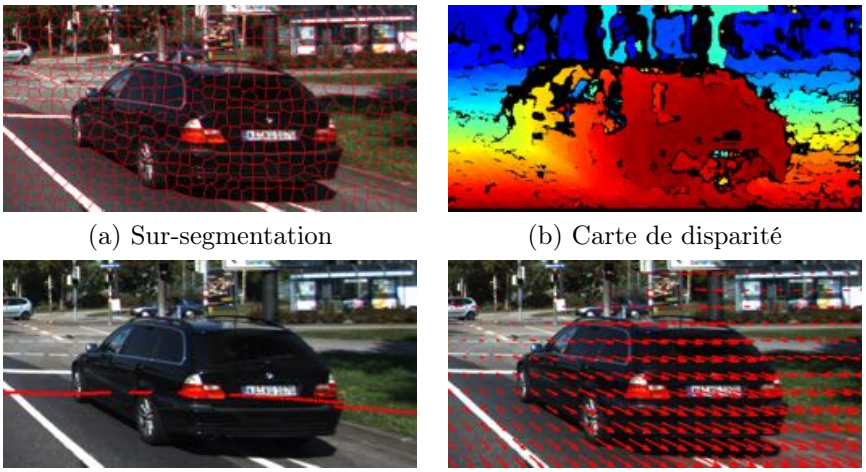
où $pl_i$ est la plausibilité associée à l'observation $X_i$. La perte $L_i = (1-pl_i(\{c_i\}))^2 + pl_i(\{\overline{c_i}\})^2$ possède les propriétés suivantes :

$$\begin{aligned} m_i(\{c_i\}) \to 1 &\Rightarrow L_i \to 0, \\ m_i(\{\overline{c_i}\}) \to 1 &\Rightarrow L_i \to 2, \\ m_i(\Omega) \to 1 &\Rightarrow L_i \to 1. \end{aligned} \tag{16}$$

La fonction de perte attribue un coût élevé pour de la masse incorrectement allouée et un coût intermédiaire pour l'ignorance. La valeur de $\beta$ est arbitrairement fixée à 2.

# 4 Application à la compréhension de scènes routières

Nous appliquons notre schéma de fusion à un système composé d'un dispositif stéréo utilisant deux caméras couleur et d'un LiDAR, que nous supposons calibré. Plusieurs modules indépendants traitent les données issues de ces capteurs afin de fournir une information sur les classes des segments de l'image. Dans nos travaux, nous nous plaçons dans le référentiel de la caméra gauche, seules les images issues de cette dernière sont classées. Les informations 3D capturées par les deux caméras et le LiDAR sont utilisées dans un premier temps



(a) Sur-segmentation (b) Carte de disparité

(c) Points laser (d) Flot optique

FIGURE 4 – Données d'entrée du système de fusion multicapteurs.

pour détecter le sol. Elles seront également utilisées, par la suite, pour la détection de tout ce qui n'est pas le ciel. Un module fondé sur l'analyse de texture est également envisagé pour la détection du ciel, de la route et de la végétation. Enfin, une propagation temporelle est utilisée afin de lier deux images consécutives. Les données d'entrée des différents modules sont représentées sur la figure 4.

## 4.1 Classification basée stéréo

La profondeur de chaque pixel d'une image acquise par une caméra peut être estimée à l'aide d'un dispositif stéréo. L'information 3D est générée par le calcul d'une carte de disparité (figure 4b), qui peut éventuellement être incomplète et erronée. Il existe de nombreuse méthodes de calcul de disparité [22]. Dans notre étude, nous avons utilisé l'approche semi-globale proposée par Hirschmüller (2008). Ce choix a été motivé par le fait que parmi les algorithmes évalués sur la base de données KITTI, l'approche de Hirschmüller (2008) est l'une des meilleures méthodes lorsque seuls les pixels dont une disparité a pu être estimée sont pris en compte. De plus, une implémentation en temps réel peut être obtenue en parallélisant les calculs sur une carte graphique [6]. Le dispositif stéréo étant calibré et ses paramètres de calibration connus, il est possible de projeter chaque pixel de la carte de disparité dans l'espace 3D. Un segment de l'image correspond alors à un nuage de points 3D.

Les informations 3D peuvent être utilisées pour détecter le sol. Pour ce faire, nous utilisons un estimateur robuste ($RANSAC$) pour détecter le plan du sol, sous l'hypothèse que le sol soit bien planaire. Des modèles plus complexes existent également dans la littérature pour traiter les cas de sol non planaire [27]. À partir de là, un segment d'image est classé comme sol ou non-sol selon sa distance par rapport au plan estimé.

Le cadre de discernement est donc $\Omega = \{Sol, \overline{Sol}\}$. Le modèle de la classe « sol » est tout simplement l'équation du plan du sol dans le référentiel de la caméra gauche, qu'on notera $\Pi$. Un segment d'image $S$ est représenté par un


(a)

(b)

FIGURE 5 – (a) Les segments rouges correspondent à ceux qui sont entièrement au-dessus de la ligne d'horizon, ils ne peuvent donc pas correspondre au ciel. Les segments verts sont sous la ligne d'horizon, ils ne peuvent donc pas correspondre au sol. (b) Les segments rouges correspondent à ceux dont la disparité moyenne est supérieure à 24 pixels. Ils ne peuvent donc pas correspondre au ciel.

ensemble de $n$ points $X = \{p_1,p_2,\ldots,p_k,p^*_{k+1},\ldots,p^*_n\}$, où les points $p^*_i$ sont ceux dont la disparité n'a pu être calculée. La distance entre l'observation $X$ et le modèle $\Pi$ est alors définie comme la distance moyenne des points valides $p_i$ au plan $\Pi$ :

$$d(X,\Pi) \quad = \quad \frac{1}{k}\sum_{i=1}^{k} d(p_i,\Pi), \tag{17}$$

où $d(p_i,\Pi)$ est la distance euclidienne entre le point $p_i$ et $\Pi$. Cette distance est alors utilisée pour construire une fonction de masse en utilisant la formule (14). Les disparités des pixels du segment $S$ pouvant ne pas tous avoir été calculées, la fonction de masse est affaiblie par un coefficient $\alpha = 1 - k/n$, représentant la proportion de points dont la disparité n'a pu être estimée par rapport au nombre total de points dans $S$. Ainsi, si la disparité d'aucun point de $S$ n'a pu être estimée, on se retrouve avec une masse vide $m(\Omega) = 1$.

Des informations supplémentaires peuvent être tirées de l'équation du plan du sol, comme par exemple la ligne d'horizon. Cette dernière aurait également pu être estimée à partir d'une seule image en passant par des méthodes d'estimation de point de fuite. La connaissance de la ligne d'horizon permet alors de dire que tout segment au dessus de celle-ci appartient forcément à la classe « non-sol ». Ce qui est représenté par une masse catégorique $m(\{\overline{Sol}\}) = 1$, qui peut être combinée avec la fonction de masse précédemment calculée en utilisant la règle de combinaison de Dempster (8).

L'analyse peut également être étendue à d'autres cadres de discernement, notamment avec $\Theta = \{Ciel, \overline{Ciel}\}$. En effet, on sait que tout segment en dessous de la ligne d'horizon ne peut appartenir à la classe « ciel ». De plus, les points du ciel pouvant être supposés à des distances presque infinies, i.e. avec une disparité

nulle, les segments de disparité moyenne strictement positive, ou plus grande qu'un certain seuil, peuvent également être classées comme « non-ciel ». Pour combiner cette nouvelle source d'information, on définit un nouveau cadre de discernement, commun à $\Omega$ et $\Theta$, $\Psi = \{Sol, Ciel, \overline{Sol \cup Ciel}\}$. Les résultats de classification issus de ces considérations sont illustrés sur la figure 5.

## 4.2 Classification basée LiDAR

Un capteur LiDAR fournit, comme une caméra stéréo, des informations 3D. La figure 4c montre les impacts lasers projetés sur l'image. Ainsi, comme dans le cas du système stéréo, un segment $S$ est perçu comme un ensemble de $k$ points 3D, $X = \{p_1, \ldots, p_k\}$, pouvant éventuellement être vide. En effet, contrairement au cas stéréo, les impacts laser ne donnent une information 3D que pour un nombre très limité de segments. Pour les segments comportant des impacts lasers, on utilise la même approche que dans le cas stéréo, en reprenant la distance (17) et la formule (14).

De plus, tout l'espace entre un impact laser et l'origine du capteur LiDAR est vide et est apparenté au sol. Au niveau de l'image, les segments $S$ par lesquels sont passés des rayons laser sont les segments se retrouvant au-dessous d'un impact laser. Ainsi, pour ces segments, on associe la fonction de masse suivante :

$$\begin{array}{rcl} m(\{Sol\}) & = & k/n \\ m(\{\overline{Sol}\}) & = & 0 \\ m(\Omega) & = & 1 - k/n \end{array} , \tag{18}$$

où $k$ est le nombre de rayons mesurés passant à travers $S$ et $n$ le nombre maximal de rayons ayant potentiellement pu traverser $S$. On peut remarquer que cette fonction de masse est, en fait, l'affaiblissement d'une masse catégorique sur $\{Sol\}$ par un coefficient $\alpha = 1 - k/n$.

## 4.3 Classification basée sur la texture

La texture est une information importante pour les tâches de classification. Pour encoder l'information de texture, la transformation de Walsh-Hadamard est utilisée, en suivant l'approche de Wojek et Schiele (2008). Les coefficients de Walsh-Hadamard se révèlent suffisants pour discriminer les classes « végétation », « ciel » et « route ». L'utilisation de caractéristiques plus complexes [29] peut être envisagée de manière similaire. Pour construire un modèle, pour une classe donnée, une approche *sac de mots* [29] est employée. Les caractéristiques de textures sont tout d'abord quantifiées en un ensemble de $N$ *textons* ; un modèle est alors simplement un histogramme normalisé $H = (h_1, \ldots, h_N)$, où $h_i$ est la fréquence d'apparition du $i^{\text{ème}}$ texton pour la classe en question. Chaque segment est également observé sous la forme d'un histogramme $X = (x_1, \ldots, x_N)$, sa distance au modèle $H$ peut alors être calculée à l'aide d'une distance entre histogrammes, comme la distance $\chi^2$ par exemple :

$$d(X,H) = \frac{1}{2} \sum_{i=1}^{N} \frac{(x_i - h_i)^2}{x_i + h_i}. \tag{19}$$

Cette approche permet de pouvoir construire un modèle pour chaque classe indépendamment les unes des autres. Seuls des exemples positifs sont nécessaires

pour apprendre le modèle. On peut alors se dispenser de considérer toutes les classes d'objets possibles, d'autant plus que pour certains types d'objets, l'information de texture n'est pas assez discriminante. Dans nos travaux, nous nous limitons aux classes « végétation », « ciel » et « route », ces trois classes étant bien adaptées à une analyse de texture.

En travaillant avec des informations de texture, il peut arriver qu'une petite distance entre une observation et un modèle ne soit pas suffisante pour inférer la classe de l'objet en question. Par exemple, la façade blanche d'un bâtiment est très proche, en terme d'apparence, du ciel, qui apparaît souvent blanc sur l'image. Toutefois, lorsque la texture est éloignée de celle du ciel, on peut déduire que l'objet n'appartient pas à la classe « ciel ». Ainsi, il est plus prudent, ici, de considérer $d^- = 0$.

### 4.4 Propagation temporelle

Enfin, un dernier module de traitement, propage le résultat de classification d'un instant $t$ à un instant $t+1$ à l'aide du flot optique (figure 4d). Ce dernier est calculé à partir de deux images consécutives en utilisant les travaux de Werlberger *et al.* (2010). À chaque segment $S_t$ de l'image à l'instant $t$ est associé un segment $S_{t+1}$ à l'instant $t+1$, défini comme celui pointé par le flot optique médian des pixels de $S_t$. La fonction de masse associée à $S_t$ est alors directement transférée à $S_{t+1}$ :

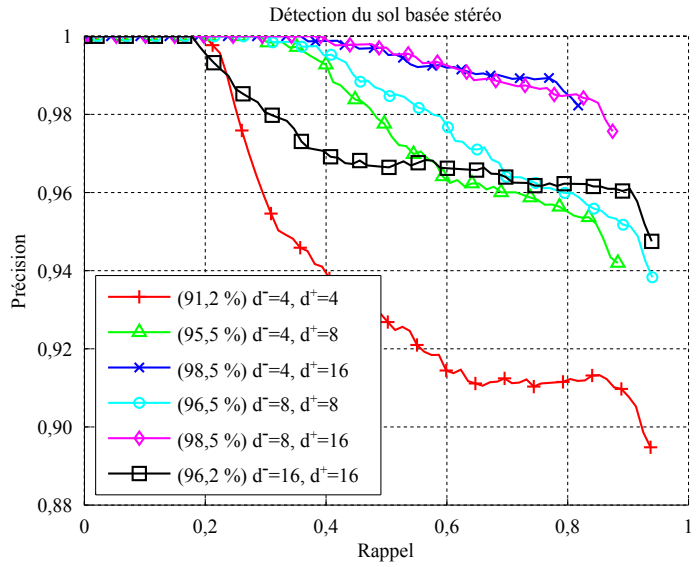
$$\forall A \subseteq \Omega, \qquad m_{t+1}(A) = m_t(A). \tag{20}$$

Elle est ensuite affaiblie par un facteur $\alpha$ correspondant au ratio de pixels de $S_t$ ne pointant pas vers $S_{t+1}$. Un segment à l'instant $t+1$ peut recouvrir plusieurs segments à l'instant $t$. Ainsi, si plusieurs segments pointent vers le même segment à $t+1$, les fonctions de masse associées sont simplement combinées par la règle de Dempster.

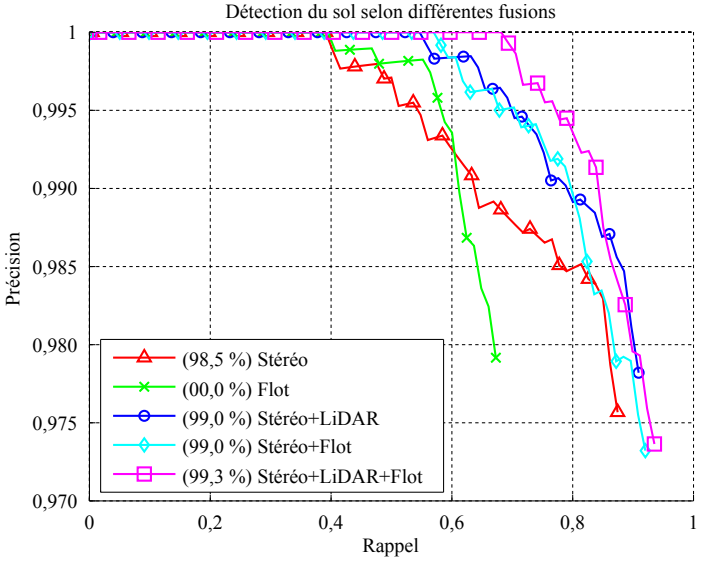
## 5 Résultats expérimentaux

La base de données KITTI [9] a été utilisée pour valider l'approche proposée. Les deux caméras couleur ainsi que le LiDAR Velodyne ont été utilisés comme capteurs. Toutefois, seule la nappe du milieu du LiDAR Velodyne a été considérée ($32^{\text{e}}$ nappe parmi les 64 nappes disponibles) afin de simuler un capteur LiDAR simple nappe, communément utilisé en robotique mobile. La calibration de ces capteurs a été faite à l'aide d'une méthode automatique utilisant plusieurs mires de calibration [10]. Un ensemble de 110 images ont été manuellement annotées à l'aide du logiciel Adobe® Photoshop® CS2, 70 images ont servi pour l'apprentissage et 40 pour le test.

Les paramètres pour chaque module ont été choisis en testant différentes configurations. La figure 6a montre l'influence des seuils $d^-$ et $d^+$ sur la détection du sol basée sur le module stéréo. Les valeurs $d^- = 8$ et $d^+ = 16$ semblent être un bon compromis entre précision et rappel. Pour un taux de rappel de 80 %, nous obtenons une précision de 98,5 %. Nous remarquons que le taux de rappel maximal est atteint pour les configurations où $d^- = d^+$. En effet, lorsque ce n'est pas le cas, il est possible de se retrouver dans le cas (13b), c'est-à-dire dans une zone d'ignorance où aucune décision ne peut être prise. En permettant de




(a)




(b)

FIGURE 6 – Courbes précision/rappel pour la détection du sol. Les valeurs entre parenthèses correspondent à la précision pour un taux de rappel de 80 %. (a) Performances du module basé stéréo pour différentes valeurs de $d^-$ et $d^+$. (b) Performances en combinant différents modules.

Classification basée sur la texture

| | Gazon | Route | Arbre | Obst | Ciel | Rappel |
|---|---|---|---|---|---|---|
| Gazon | | | | | | 0 |
| Route | | 66 | | 33,8 | 0,2 | 50,2 |
| Arbre | | | | | | 0 |
| Obst | | 14,3 | | 84,1 | 1,6 | 49,3 |
| Ciel | | 0 | | 18,4 | 81,6 | 80,5 |

(a)

Classification après fusion

| | Gazon | Route | Arbre | Obst | Ciel | Rappel |
|---|---|---|---|---|---|---|
| Gazon | 81,1 | 3,8 | 15,1 | 0 | 0 | 40,6 |
| Route | 7,9 | 89,1 | 0,3 | 2,7 | 0 | 78,8 |
| Arbre | 2,5 | 0 | 94,4 | 3,1 | 0 | 86,7 |
| Obst | 0,8 | 2,3 | 9,1 | 86,8 | 1 | 52,6 |
| Ciel | 0 | 0 | 0 | 18,4 | 81,6 | 80,5 |

(b)

FIGURE 7 – Matrices de confusion pour la classification multiclasse. (a) Résultats de la classification basée uniquement sur l'analyse de texture. (b) Résultats après fusion avec les modules stéréo, LiDAR et flot optique.

ne pas prendre de décision dans les cas ambiguës, la précision est grandement améliorée. Quant aux paramètres $\beta$ et $\gamma$, ils n'ont pas d'influence sur la précision ni le rappel lorsque le module est considéré tout seul. $\beta$ est arbitrairement fixé à 2 tandis que $\gamma$ est choisi par rapport à la fonction de perte (15).

En combinant plusieurs détecteurs de sol, issus de la stéréo, du LiDAR et du flot optique, les performances sont clairement améliorées comme illustré sur la figure 6b. Nous remarquons que la combinaison de plusieurs sources est toujours meilleure que chacune des sources prise individuellement. Les figures 8(b-e) illustrent les différentes masses affectées aux classes « sol » et « non-sol ».

Le module fondé sur l'analyse de texture a été utilisé pour la détection de la route, de la végétation et du ciel. La figure 7a montre la matrice de confusion de la classification multiclasse et les figures 9(a-c) illustrent les résultats obtenus. Utilisé seul, ce module ne peut différencier le gazon des arbres. En revanche, cela devient possible en le combinant avec les détecteurs de sol. La figure 7b montre la matrice de confusion du système complet. Nous voyons encore une fois une amélioration des résultats. En particulier, un gain de plus de 20 % en précision et en rappel a été obtenu sur la classe « route ». En revanche, comme les détecteurs de sol ne donnent aucune information sur le ciel, les résultats pour la classe « ciel » sont inchangés.

Nous rappelons que, dans certains cas, aucune décision ne peut être prise (cas de l'ignorance), ce qui explique que le taux de rappel soit inférieur à la diagonale de la matrice de confusion. Nous pouvons, par exemple, remarquer que le taux de rappel de la classe « gazon » est particulièrement bas. Cela s'explique par le fait que le gazon est parfois légèrement surélevé par rapport au sol. Le gazon n'est alors même pas classé comme sol, comme nous pouvons le voir sur la partie en bas à gauche de l'image 9d où les régions non colorées sont celles n'ayant pas été classées. Cela montre également que la définition de la classe « gazon » comme étant l'intersection entre les classes « sol » et « végétation » n'est pas absolument correcte.

La figure 10 montre plusieurs résultats obtenus après la fusion de tous les modules. L'analyse n'étant faite qu'à un niveau très local, il existe de nombreux

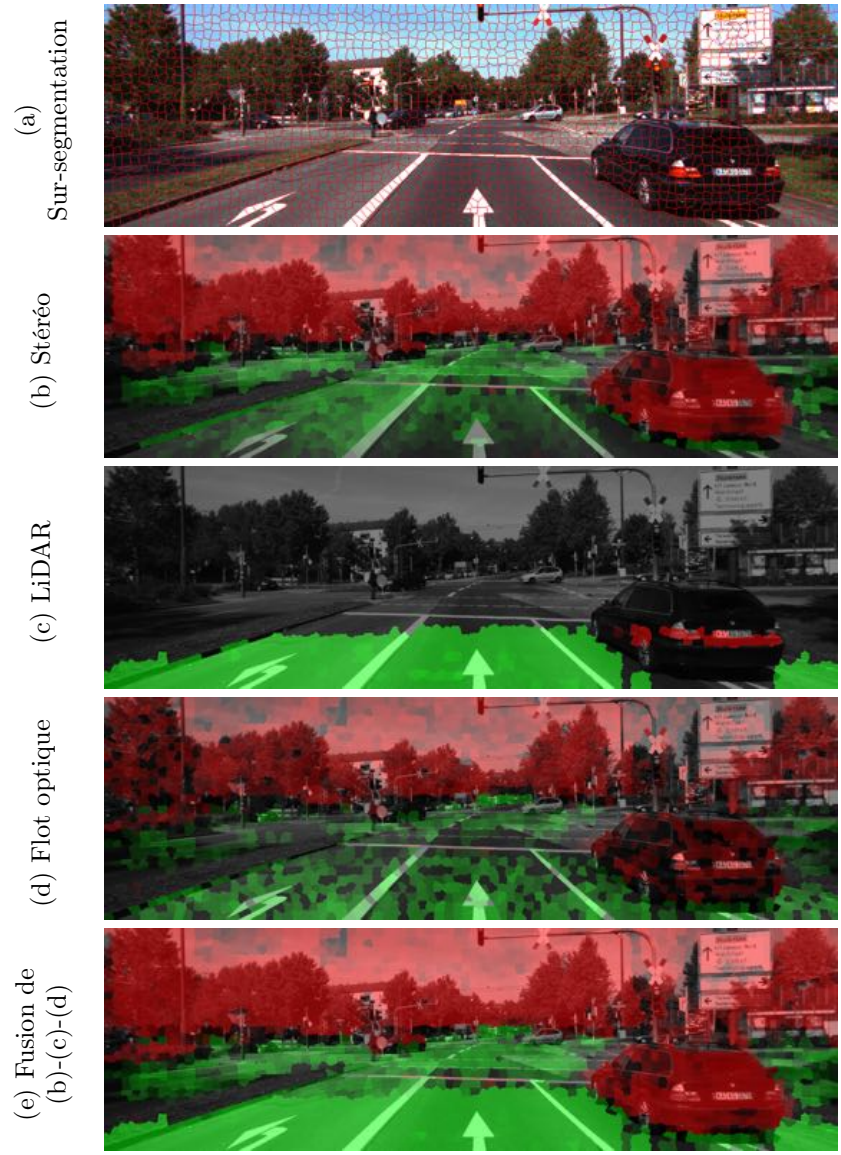



FIGURE 8 – (a) Image sur-segmentée. (b-e) Résultats de détection du sol avec $\Omega = \{Sol, \overline{Sol}\}$. L'intensité en vert représente la masse affectée à $\{Sol\}$ tandis que le rouge représente la masse sur $\{\overline{Sol}\}$. En l'absence de couleur, la masse est allouée à $\Omega$.

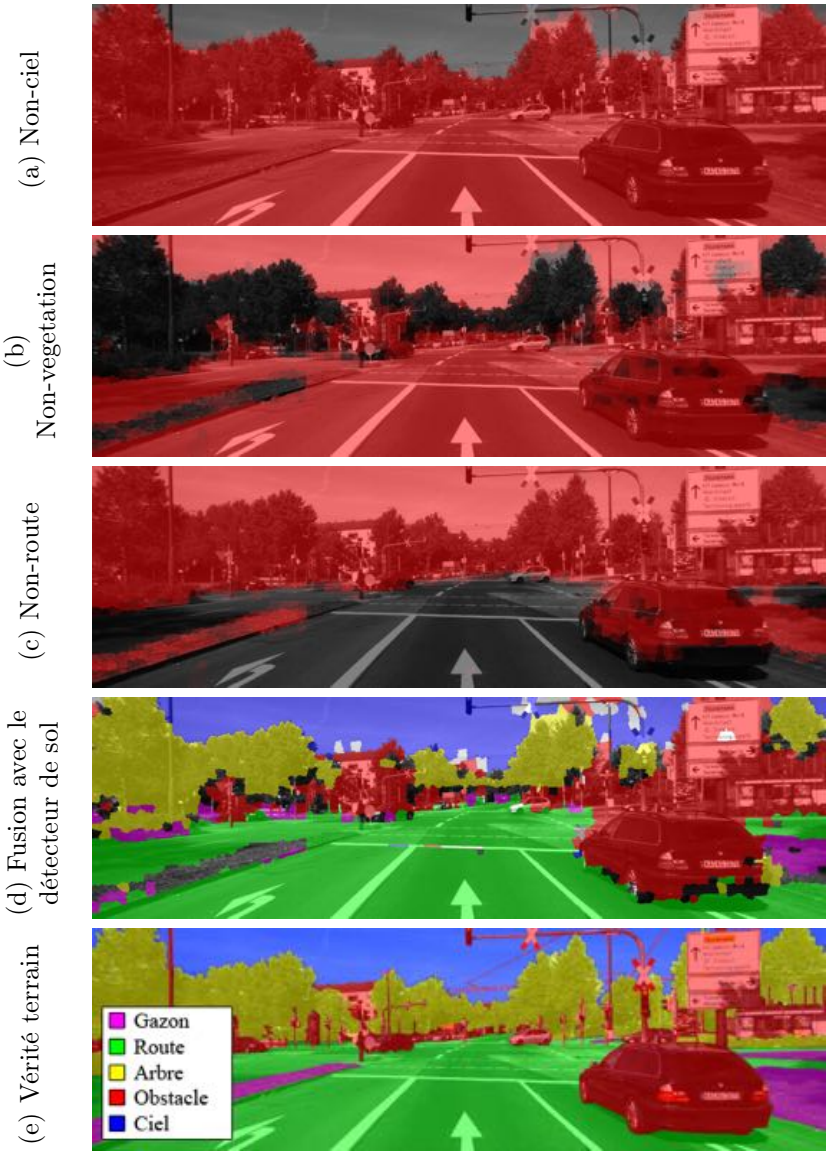



FIGURE 9 – (a-c) Analyse par texture des classes ciel, végétation et route, les masses sont uniquement allouées aux classes complémentaires (représentées en rouge). (d) Fusion de tous les modules, avec la détection de sol, le code couleur est celui de la vérité terrain (e).

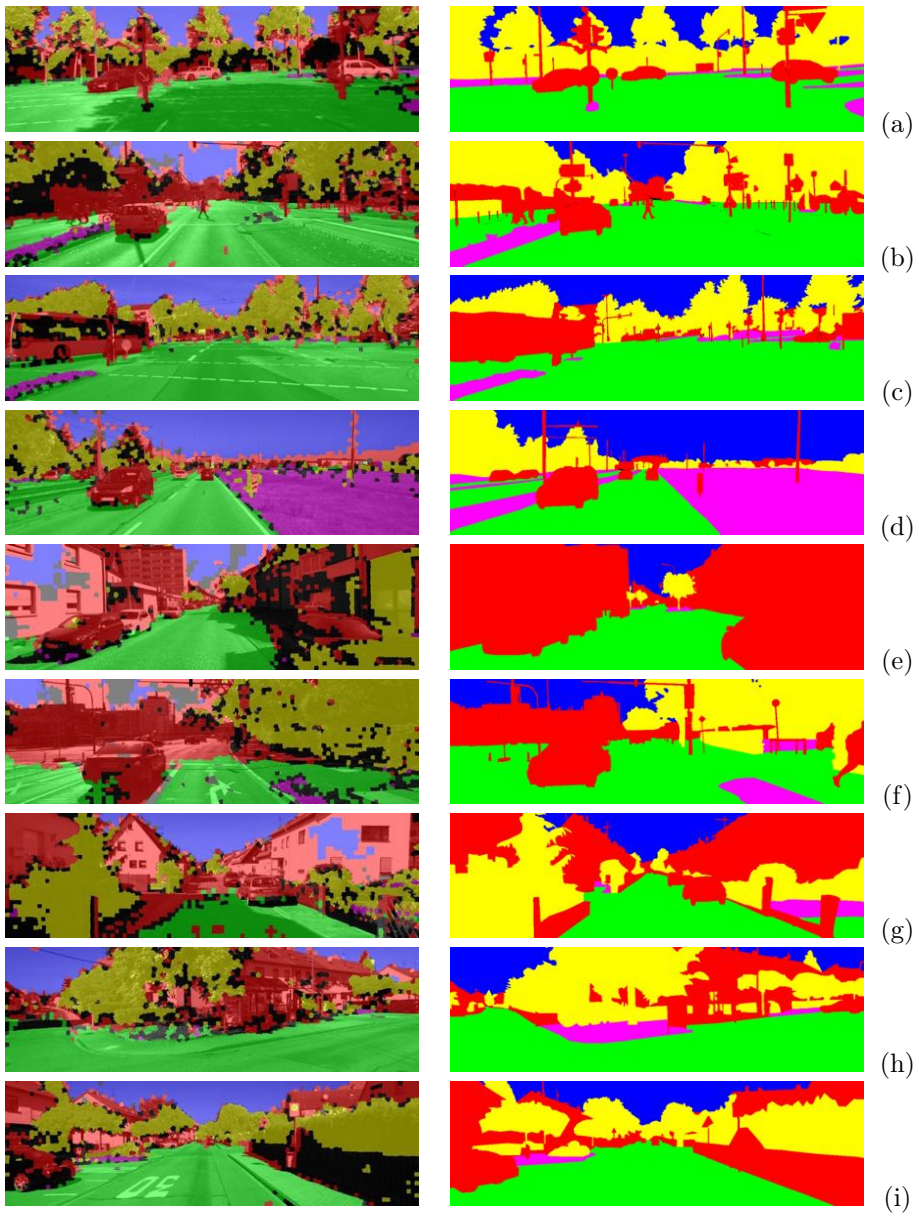



FIGURE 10 – (Gauche) Résultats obtenus après la fusion de tous les modules. (Droite) Vérités terrains avec le même code couleur que sur la figure 9e.

trous où aucune décision n'est prise. Une approche plus globale en incluant des contraintes entre segments adjacents fera l'objet de travaux futurs et pourra combler certains trous. De même, l'ajout de nouvelles sources d'information pourra corriger certaines erreurs et aider à la décision. Nous pouvons également remarquer des erreurs dues à la sur-segmentation, en particulier au niveau des structures fines comme les poteaux (figure 10(d,f)). La contrainte de régularité en taille des segments imposée par l'algorithme SLIC ne permet pas de correctement segmenter les objets de petite taille. L'utilisation de plusieurs algorithmes de sur-segmentation comme dans [17] pourrait partiellement résoudre ce problème.

Le calcul de la sur-segmentation, de la disparité et du flot optique peuvent être fait de manière parallèle. Il existe, dans la littérature [20, 6, 28], plusieurs implémentations en temps réel de ces algorithmes, qui n'ont toutefois pas été considérées dans ces travaux. Puisque tous les modules traitent les données de manière indépendante, ils peuvent être lancés en parallèle. Les traitements de chaque module sont tous relativement simples et rapides à exécuter. Enfin le coût de la combinaison des masses qui peut être, en théorie, très élevé ne l'est pas en pratique car seules les plausibilités sur les singletons sont considérées.

# 6 Conclusions et perspectives

Nous avons proposé un schéma original de fusion d'informations fondé sur une sur-segmentation et la théorie des fonctions de croyance. Il est suffisamment flexible pour pouvoir rajouter de nouvelles classes d'objets, de nouveaux modules de traitement et de nouveaux capteurs. Cette flexibilité vient notamment du fait que les modules peuvent fonctionner indépendamment les uns des autres. Ce choix peut cependant avoir des répercussions sur les performances globales. En effet, il peut sembler plus judicieux, étant donné un certain nombre de données et de modules, d'apprendre un modèle global ou une re-pondération optimale de ces modules. Toutefois, dans un contexte où la flexibilité et la robustesse priment sur les performances, il est souhaitable de ne pas à avoir à entraîner le système entier à chaque ajout de modules.

Des travaux futurs seront menés afin d'ajouter de nouvelles classes comme les piétons ou les véhicules, en adaptant notamment des méthodes fondées sur l'utilisation de fenêtres glissantes. En combinant avec une information de profondeur, l'information au niveau d'une boîte englobante pourra, par exemple, être ramenée à celui des segments de l'image. De nouvelles sources d'information comme le GPS ou les cartes seront également considérées pour la détection d'objets en mouvement. Enfin, une approche globale sera également étudiée afin de fusionner des segments voisins appartenant au même objet, ce qui permettra d'avoir une compréhension à plus haut niveau de la scène.

# Remerciements

Ce travail mené dans le cadre du Labex MS2T s'insère dans le programme Investissements d'Avenir géré par l'Agence Nationale de la Recherche (ANR-11-IDEX-0004-02). Il est soutenu et financé par le projet 26193PE du programme Cai Yuanpei, accordé par le ministère chinois de l'Éducation et les ministères

français des Affaires Étrangères et Européennes et de l'Enseignement Supérieur et de la Recherche, ainsi que par le projet Blanc International ANR-NSFC franco-chinois PRETIV (ANR-11-IS03-0001).

# Références

[1] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Susstrunk. SLIC superpixels compared to state-of-the-art superpixel methods. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 34(11) :2274–2282, 2012.

[2] H. Badino, U. Franke, and R. Mester. Free space computation using stochastic occupancy grids and dynamic programming. In *Proc. International Conference on Computer Vision Workshop on Dynamical Vision*, Rio de Janeiro, Brazil, 2007.

[3] J. A. Barnett. Calculating Dempster-Shafer plausibility. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 13(6) :599–602, 1991.

[4] T. Denœux. A k-nearest neighbor classification rule based on Dempster-Shafer theory. *IEEE Trans. on Systems, Man and Cybernetics*, 25(5) :804–813, 1995.

[5] P. Dollár, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection : an evaluation of the state of the art. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 34(4) :743–761, 2011.

[6] I. Ernst and H. Hirschmüller. Mutual information based semi-global stereo matching on the GPU. In *Proc. International Symposium on Advances in Visual Computing*, pages 228–239, Las Vegas, USA, 2008.

[7] A. Ess, T. Müller, H. Grabner, and L. Van Gool. Segmentation based urban traffic scene understanding. In *Proc. British Machine Vision Conference*, pages 1–11, London, UK, 2009.

[8] C. Farabet, C. Couprie, L. Najman, and Y. LeCun. Scene parsing with multiscale feature learning, purity trees, and optimal covers. In *Proc. Internation Conference on Machine Learning*, Edinburgh, Scotland, 2012.

[9] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving ? The KITTI vision benchmark suite. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pages 3354–3361, Providence, USA, 2012.

[10] A. Geiger, F. Moosmann, O. Car, and B. Schuster. Automatic camera and range sensor calibration using a single shot. In *Proc. IEEE International Conference on Robotics and Automation*, pages 3936–3943, Saint Paul, USA, 2012.

[11] A. Geiger, C. Wojek, and R. Urtasun. Joint 3D estimation of objects and scene layout. In *Proc. Conf. on Neural Information Processing Systems*, pages 1467–1475, Granada, Spain, 2011.

[12] D. Hoiem, A. A. Efros, and M. Hebert. Recovering surface layout from an image. *International Journal of Computer Vision*, 75(1) :151–172, 2007.

[13] D. Hoiem, A. A. Efros, and M. Hebert. Closing the loop on scene interpretation. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pages 1–8, Anchorage, USA, 2008.

[14] L. Ladický, P. Sturgess, C. Russell, S. Sengupta, Y. Bastanlar, and *et al.* Joint optimization for object class segmentation and dense stereo reconstruction. *International Journal of Computer Vision*, 100(2) :122–133, 2012.

[15] B. Leibe, N. Cornelis, K. Cornelis, and L. Van Gool. Dynamic 3D scene analysis from a moving vehicle. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pages 1–8, Minneapolis, USA, 2007.

[16] A. Levinshtein, A. Stere, K. N. Kutulakos, D. J. Fleet, S. J. Dickinson, and K. Siddiqi. TurboPixels : Fast superpixels using geometric flows. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 31(12) :2290–2297, 2009.

[17] S. Mathevet, L. Trassoudaine, P. Checchin, and J. Alizon. Combinaison de segmentations en régions. *Traitement du Signal*, 16(2) :93–104, 1999.

[18] J. Moras, V. Cherfaoui, and P. Bonnifait. Moving objects detection by conflict analysis in evidential grids. In *Proc. IEEE Intelligent Vehicles Symposium*, pages 1120–1125, Baden-Baden, Germany, 2011.

[19] A. P. Morre, S. J. D. Prince, J. Warrel, U. Mohammed, and G. Jones. Scene shape priors for superpixel segmentation. In *Proc. IEEE International Conference on Computer Vision*, pages 771–778, Kyoto, Japan, 2009.

[20] C. Y. Ren and I. Reid. gSLIC : a real-time implementation of SLIC superpixel segmentation. Technical report, University of Oxford, Department of Engineering Science, 2011.

[21] S. A. Rodríguez, V. Frémont, P. Bonnifait, and V. Cherfaoui. Multi-modal object detection and localization for high integrity driving assistance. *Machine Vision and Applications*, 14 :1–16, 2011.

[22] D. Scharstein and R. Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International Journal of Computer Vision*, 47(1–3) :7–42, 2002.

[23] G. Shafer. *A mathematical theory of evidence*. Princeton University Press, Princeton, New Jersey, 1976.

[24] P. Smets and R. Kennes. The transferable belief model. *Artificial Intelligence*, 66 :191–243, 1994.

[25] S. Thrun, W. Burgard, and D. Fox. *Probabilistic robotics (Intelligent Robotics and Autonomous Agents)*. The MIT Press, Cambridge, Massachusetts, 2005.

[26] C. C. Wang, C. Thorpe, S. Thrun, M. Hebert, and H. Durrant-Whyte. Simultaneous localization, mapping and moving object tracking. *International Journal of Robotics Research*, 26(1) :889–916, 2007.

[27] A. Wedel, H. Badino, C. Rabe, H. Loose, U. Franke, and D. Cremers. B-spline modeling of road surfaces with an application to free-space estimation. *IEEE Trans. on Intelligent Transportation Systems*, 10(4) :572–583, 2009.

[28] M. Werlberger. *Convex Approaches for High Performance Video Processing*. PhD thesis, Graz University of Technology, 2012.

[29] J. Zhang, M. Marszalek, S. Lazebnik, and C. Schmid. Local features and kernels for classification of texture and object categories : a comprehensive study. *International Journal of Computer Vision*, 73(2) :213–238, 2007.